

PATENT APPLICATION
SYSTEM AND METHOD FOR DETECTING
GENERALIZED SPACE-TIME CLUSTERS

5 This invention was developed under Government Contract
(DARPA Project F30602-01-C-0184). The U.S. Government has
rights to this patent application.

BACKGROUND OF THE INVENTION

10 1. Field of Invention:

The present invention relates to modeling space-time
data, and more particularly to detecting three-dimensional
convex clusters in space and time for a phenomenon.

15 2. Discussion of Related Art:

Detection of clusters in space and time, called
space-time clusters, is an important function in various
domains. For example, detection of such clusters is an
important part of the investigation of disease outbreaks in
20 the domain of epidemiology and public health. Other domains
of application include medical imaging, urban planning and
reconnaissance. The notion of what constitutes a cluster
depends on the domain. For example, the spatial scan
statistic as described in the paper "A spatial scan

statistic" by Martin Kulldorff in Communications in
Statistics: Theory and Methods, Volume 26, Number 6, 1997,
is widely used in the epidemiology and public health domain.

Other models of clustering might be appropriate in other
domains.

Methods for detecting clusters may be developed
depending on the clustering notion used. For example, the
use of the scan statistic implies that earlier hierarchical
approaches (see for example, "Automatic subspace clustering
of high dimensional data for data mining applications" by R.
Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan in
Proceedings of the ACM-SIGMOD International Conference on
Management of Data, 1998) to clustering cannot be applied.
An example of a system that may handle the spatial scan
statistic model for clustering is the SaTScan system.
SaTScan may be used to detect space-time clusters with a
cylindrical shape, representing a circular region in space
for the entire duration of an interval in time.

SUMMARY OF THE INVENTION

The cylindrical shape may not represent clusters that
shrink or grow with time. Also, it may not represent
movement of the phenomenon over time. The exhaustive search
based on a grid that is utilized by systems like SaTScan may
not be extended to more general shapes. Therefore, a need

exists for a system that at least may detect generalized space-time clusters that model such characteristics of the underlying phenomenon.

A system for detecting clusters in space and time using input data on occurrences of a phenomenon and characteristics at a plurality of locations and times includes an expectation generation module, an occurrence modeling module, a search module, a convex container module, and a solution evaluation module.

The expectation generation module determines expected occurrences of a phenomena at a plurality of locations and a plurality of times, and the occurrence modeling module determines actual occurrences of the phenomena at a plurality of locations and a plurality of times. The search module searches the expected occurrences and the actual occurrences for a plurality of candidate solutions. Each solution is represented as a set of points in the three-dimensional space, wherein each point corresponds to a location at a time. The convex container module determines at least one solution corresponding to a selected convex container shape from the plurality of candidate solutions, and the solution evaluation module determines a strength metric for each solution determined by the convex container module, the search module selecting a solution having a desirable strength, wherein the solution having the

desirable strength indicates a dominant cluster in the input data.

The search module selects a strongest solution as determined by the solution evaluation module.

5 A cache module may be included to save the solutions having the desired shape determined by the convex container module for previously examined sets of points.

10 The input data on occurrences of a phenomenon include counts and times of the occurrences of the phenomenon at the locations in a time period. The input data on characteristics of the locations and times include the populations subject to the occurrences of the phenomenon at the locations and times.

15 The expectation generation model generates expected counts of occurrences at the locations and times using a Poisson model.

20 The occurrence modeling module determines the occurrences as equal to the occurrences in the input data. The occurrence modeling module determines the occurrences at the locations and times based on their characteristics and a domain dependent model. The occurrences are determined from the population using a Poisson model.

25 The search module considers candidate solutions represented as sets of points and utilizes the convex container module to determine solutions having the desired

shape from the candidate solutions. The candidate solutions are initialized based on the input data. Each initial candidate solution is singleton point.

The search module determines candidate solutions from solutions considered using an iterative process. Candidate solutions are created from solutions considered based on the chosen convex container shape.

The convex container module may determine a solution in various ways. For example, the convex container module determines a solution with minimum volume, given the selected convex container shape, that includes all the points in a given candidate solution. The convex container module may determine a solution with maximum volume, given the selected convex container shape, that excludes all the points not in the given candidate solution. According to another example, the convex container module may determine a solution that maximizes a measure representing the equality between the set of points in the given candidate solution and the set of points included in the solution, given the selected convex container shape.

The solution evaluation module determines the strength metric based on all the points included in the solution and the expected occurrences determined by the expectation generation module and the occurrences determined by the occurrence modeling module.

The strength metric is based on the likelihood ratio using the spatial scan statistic.

BRIEF DESCRIPTION OF THE DRAWINGS

5 Preferred embodiments of the present invention will be described below in more detail, with references to the accompanying drawings:

Figure 1 is an illustration of a system according to an embodiment of the present disclosure;

10 Figure 2 is an illustration of a system according to an embodiment of the present disclosure;

Figures 3A and 3B are illustrations of a cluster with a truncated square pyramid shape according to an embodiment of the present disclosure where the three dimensional view is given in Figure 3A and the two dimensional projection from the top on to the spatial plane is given in Figure 3B;

Figure 4 is an illustration of a cluster with a cylindrical shape;

20 Figures 5A and 5B are illustrations of a cluster with a truncated cone shape according to an embodiment of the present disclosure where the three dimensional view is given in Figure 5A and the two dimensional projection from the top on to the spatial plane is given in Figure 5B;

Figure 6 shows results according to an embodiment of the present disclosure for cancer data using a square pyramid as a selected convex container shape; and

Figure 7 is a flow chart of a method according to an embodiment of the present disclosure.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

The detection of three-dimensional clusters in space and time is used for determining localized occurrences of a phenomenon of interest. Data comprising space and time information is analyzed to determine where and when the phenomenon of interest appears. A system and/or method analyze the input data to determine a phenomenon of interest and use a user selected three dimensional shape to determine properties of the phenomenon.

It is to be understood that the present invention may be implemented in various forms of hardware, software, firmware, special purpose processors, or a combination thereof. In one embodiment, the present invention may be implemented in software as an application program tangibly embodied on a program storage device. The application program may be uploaded to, and executed by, a machine comprising any suitable architecture.

Referring to Fig. 1, according to an embodiment of the present disclosure, a computer system 101 for implementing

the present invention may comprise, *inter alia*, a central processing unit (CPU) 102, a memory 103, and an input/output (I/O) interface 104. The computer system 101 is generally coupled through the I/O interface 104 to a display 105 and various input devices 106 such as a mouse and keyboard. The support circuits may include circuits such as cache, power supplies, clock circuits, and a communications bus. The memory 103 may include random access memory (RAM), read only memory (ROM), disk drive, tape drive, etc., or a combination thereof. The present invention may be implemented as a routine 107 that is stored in memory 103 and executed by the CPU 102 to process the signal from the signal source 108. As such, the computer system 101 is a general-purpose computer system that becomes a specific purpose computer system when executing the routine 107 of the present invention.

The computer platform 101 also includes an operating system and microinstruction code. The various processes and functions described herein may either be part of the microinstruction code or part of the application program (or a combination thereof), which is executed via the operating system. In addition, various other peripheral devices may be connected to the computer platform such as an additional data storage device and a printing device.

It is to be further understood that, because some of the constituent system components and method steps depicted in the accompanying figures may be implemented in software, the actual connections between the system components (or the process steps) may differ depending upon the manner in which the present invention is programmed. Given the teachings of the present disclosure provided herein, one of ordinary skill in the related art will be able to contemplate these and similar implementations or configurations of the present invention.

Referring to Figure 2, a system and method for detecting space-time clusters is depicted. The clusters detected by this system may model growth or shrinkage over time of the phenomenon of interest. The clusters may also model movement of the phenomenon of interest with time. The input data 201 comprises information on the occurrences of the phenomenon being analyzed by locations and times 202, and the characteristics of these locations and times 203. The system is applicable in various domains. The information on the occurrences of the phenomenon 202 is the counts of the occurrences for each combination of location and time for the region and time period of interest. For example, in the public health domain, the information on the occurrences of the phenomenon 202 may comprise counts of occurrences of a particular disease for each county and

month/year. The characteristics of these locations and times
203 comprises population counts that are subject to the
phenomenon at each combination of location and time. For
example, the characteristics of these locations and times
5 203 may comprise the population of the counties being
considered at various times in the period of interest along
with other relevant demographic information on the
population (e.g., age, gender, etc.). A selected convex
container shape for the detected clusters determines the
10 properties of the phenomenon that may be modeled.

Referring to Figure 2, the expectation generation
module 204 determines the expected occurrences at the
locations and times using the input data. The expected
values are determined using the input characteristics of
15 locations and times 203 and a domain dependent model for the
expected behavior in the absence of any clustering. For
example, in the epidemiology domain the expected counts for
the occurrences of a disease may be modeled in the absence
of any clustering using the Poisson distribution based on
20 the population. Consider the analysis of the occurrences of
a disease in a set of counties over an interval of a time.
The expected disease counts for any county in any year may
be determined as being proportional to the population of the
county during that year. Consider the example where the
25 total number of disease cases in 10 counties over 10 years

is 1000 and the total population is 100000 over all the 10 counties in each of the 10 years. The expected number of cases in a county with a population of 20000 in one year is 20. Other characteristics of the population may also be available. Demographics of the population like age and gender may be factored into the expectation generation by well-known statistical methods like indirect standardization.

Referring to Figure 2, the occurrence modeling module 205 determines the occurrences at each location and time. The occurrence modeling module 205 operates in different modes. A first mode is used to determine a strongest cluster in the input data of occurrences, where the strength of the cluster is determined by the solution evaluation module 208. In the first mode, the occurrence modeling module 205 uses the occurrences of the phenomenon 202 of the input data for analysis. Occurrence counts for the phenomenon are determined by location and time. A second mode is used to determine the strongest cluster of occurrences in a simulated environment. The model used as the basis for the expectation generation module 204 (e.g., a random sampling using a Poisson distribution) is used to determine the occurrences in the simulated environment. In the second mode, the occurrence modeling module 205 determines the occurrences based on the characteristics of the locations

and times provided in the characteristics of these locations and times 203 of the input data using the model for expected occurrences in the absence of any clustering. For example, the occurrences are determined by random sampling from a Poisson distribution where the expected counts are proportional to the population information in the characteristics of these locations and times 203 of the input data. Other information, such as demographics in the characteristics of these locations and times 203 of the input data may be factored in using modeling methods including, for example, indirect standardization. In the second mode the clusters in the random data of occurrences are determined. Experiments with these random data are used to determine the likelihood of finding clusters by chance that are comparable in strength to the strongest cluster detected in the actual data of occurrences. The cluster detected in the actual data of occurrences is significant if such strong clusters are unlikely to exist in the random datasets.

Referring to Figure 2, solutions in the search module 206 are represented by a set of points, where each point corresponds to a specific location at a specific time. The search module 206 calls the convex container module 207, wherein the convex container module 207 generates a solution corresponding to a user selected convex container shape from

any given candidate solution containing a set of points. A solution of the convex container module 207 exists if a container with the selected convex container shape may be determined that includes exactly the points in the solution.

5 The convex container module 207 makes a list of convex container shapes available to the user. The user selects from this list the convex container shape that models the phenomenon of interest most closely as determined by the user.

10 A solution of the convex container module 207 may be generated in any of a plurality of methods according to an embodiment of the present invention. For example, the convex container module generates a solution with minimum volume that includes all the points in the given set.

15 According to another example, the convex container module generates a solution with maximum volume that excludes all the points not in the given set. According to yet another example, the convex container module generates a solution that is closest to the given set of points using a measure
20 of equality between the given set of points and the set of points included in the solution. A measure of equality between two sets is the number of items that are only in one of the sets and minimizing this number would make the two sets closer to each other.

A cache module may be added to the cluster detection system for efficiency purposes. The cache module stores the solutions generated by the convex container module 207 for previously encountered sets of points. If a set of points is encountered more than once, the solution for it may be gotten from the cache module instead of invoking the convex container module again.

Convex container shapes may take any of a plurality of shapes. For example, a truncated pyramid with a square cross-section, where the square cross-section at any time in an interval represents the spatial extent of the cluster at the time. An example of a three-dimensional square pyramid cluster is shown in Figures 3A and 3B. Figure 3A shows the three-dimensional view and Figure 3B shows the boundaries of the square pyramid cluster at discrete times in the projection from the top on to the spatial plane. Contrast the flexibility of the square pyramid cluster of Figure 3A with the cylindrical shape in the prior art shown in Figure 4. The square pyramid shape of Figure 3A may model growth or shrinkage over time. Since the pyramid's axis along the time dimension is not restricted to be orthogonal to the two spatial axes it may also model movement over time.

Such a square pyramid may be specified by eight parameters as follows. Referring to Figure 3A the first two parameters are the earliest time 301 and latest time 302 in

the pyramid interval. Referring to Figure 3B the next two parameters, A and B, represent the two spatial coordinates of the anchor vertex of the square cross-section 303 at the earliest time in the pyramid interval. The vertex of the cross-section with the smallest values for the spatial coordinates is chosen as the anchor vertex. Referring to Figure 3B, the parameter G is the side of the square cross-section 303 at the earliest time in the pyramid interval. Similarly, the parameters C and D represent the spatial coordinates of the anchor vertex and the parameter H is the side of the square cross-section 304 at the latest time in the pyramid interval. The volume of the square pyramid is proportional to $(G*G+G*H+H*H)$.

Accordingly, a solution of the convex container module 207 may be generated from a given set of points by solving a quadratic programming problem in which the volume as specified above using the parameters G and H of the square pyramid is minimized. Each point in the given set leads to four linear constraints in this quadratic programming formulation. These four constraints specify that the point is contained in the cross-section of the square pyramid at the specific time corresponding to the point. The cross-section of the square pyramid may be determined from the six parameters, A,B,C,D,G and H, using linear interpolation. The quadratic programming problem may be

solved using any of a plurality of methods. For example, a quadratic programming problem with linear constraints may be solved exactly using any quadratic programming library package. According to another example, the quadratic programming problem may be solved approximately by considering a set of square pyramids that contain all the points in the given set and choosing the one with minimum volume. The square pyramid shape is one example of the convex container shapes that may be selected by the user for the convex container module 207.

Another example of a convex container shape is a truncated pyramid with a regular polygon for its cross-section. The generation of a solution by the convex container module 207 from a given set of points for this truncated pyramid container shape is an extension of the determination described above for the square pyramid. The volume of this pyramid is a similar function of the sides of the regular polygon at the minimum and maximum times of the pyramid interval. Each point in the given set leads to one linear constraint for each side of the polygon. The quadratic programming problem may be solved as described above for the square pyramid case. The truncated pyramid with a regular polygon is another example of the convex container shapes that may be selected by the user.

Another convex container shape that may be selected is a truncated cone with a circular spatial cross-section for each time in an interval. An example of this truncated cone is shown in Figures 5A and 5B. Figure 5A shows the three-dimensional view and Figure 5B shows the projection of the cones outline at discrete points in time on the spatial plane. The centers of the circular cross-sections of the cone are also marked (*) in Figure 5B. A truncated cone is specified by eight parameters. Referring to Figure 5A, the first two parameters 501 and 502 are the earliest and latest time for the cone's time interval. Referring to Figure 5B, the next two parameters 503 and 504 are the spatial coordinates of the center of the circular cross-section at the earliest time in the cone's interval. Parameters 505 and 506 are the spatial coordinates of the center of the circular cross-section at the latest time in the cone's interval. Parameter 507 and 508 are the radii of the cross-sections at the earliest and latest times in the cone's interval, respectively. A solution of the convex container module 207 may be generated from the given set of points for this convex container shape using a non-linear optimization formulation. This formulation minimizes the volume that is a quadratic function of the two parameters representing the radii of the cross-sections at the earliest and latest times in the cone's interval. Each point in the

given set leads to a non-linear constraint that specifies that it is contained within the cross-section of the cone at the specific time corresponding to the point. The constraint is easily derived by simple geometric equations for the circle once the center and the radius of the cross-section at the time corresponding to the point is determined using linear interpolation. This non-linear optimization is solved by approximate methods that consider only a subset of cones containing all the points in the given set and selecting the one with the smallest volume. The truncated cone is one of the convex container shapes that may be selected by the user.

One having ordinary skill in the art would recognize, in light of the present disclosure, that other convex containers may be used for clustering.

Referring to Figure 2, the search module 206 searches for a strongest solution for the occurrences and expectations from modules 204 and 205 respectively, and the chosen convex container shape. The strength of a solution of the convex container module 207 is determined by the solution evaluation module 208. The search module 206 considers candidate solutions that are represented as sets of points and utilizes the convex container module 207 to generate solutions from these candidate solutions. The search module 206 creates candidate solutions in an

iterative process from solutions determined by the convex container module 207 considered earlier in the iterative process. This iterative process may be initialized to a set of candidate solutions containing single points that have
5 non-zero occurrences. The search module 206 stores a set of solutions determined by the convex container module 207 considered earlier in the iterative process with preference to storing the stronger solutions. Candidate solutions are created from a pair of stored solutions (P, Q) by first
10 splitting the solutions P and Q using a random three-dimensional hyperplane and combining the pieces to create new sets of points as candidates. P and Q are chosen in a biased random fashion giving preference to the stronger solutions in the stored set of solutions. Candidate
15 solutions are also created from a single stored solution R by modifying it based on the characteristics of the convex container shape corresponding to R. One or more points not in R but close to the convex container boundaries of the solution R are added to R to create a candidate solution.
20 Similarly, one or more points in R and close to the convex container boundaries of the solution R are removed from R to create a candidate solution. The occurrence information at locations and times may be used in the selection of points for addition or removal. The strongest solution determined
25 by the solution evaluation module 208 among those solutions

determined by the convex container module 207 found in the search is output as a detected cluster.

A heuristic model may be used for determining sets of points likely to contain a strongest solution fitting a convex container shape. Therefore, the search module may consider a subset of all possible sets of points while having a measure of confidence that a strongest solution fitting a convex container shape will be determined.

The search module 206 may select candidate solutions using other methods, for example, by systematically considering all candidate solutions based on a repetitive process such as a grid. This does not imply that all solutions will be explored for the chosen convex container shape. This approach may be effective for some simple shapes.

Referring to Figure 2, the solution evaluation module 208 determines a strength for any given solution of the convex container module 207. The strength measure may be customized based on the application domain. The strength measure is based on the points included in the solution and on the occurrences and expectations for the locations and times. One example of a strength measure is the likelihood ratio based on the spatial scan statistic as described in the paper "A spatial scan statistic" by Martin Kulldorff in Communications in Statistics: Theory and Methods, Volume 26,

Number 6, 1997. Another example of a strength measure is simply the number of occurrences at the points included in the solution. Consider the case where input data includes occurrences of a phenomenon at various locations and time and also the corresponding population sizes. Another example of a strength measure is the occurrence density as defined by the total number of occurrences at the points included in the solution divided by the total population corresponding to the points included in the solution. The density could also be modified to take into account the spatial extent of the solution in addition to the population. Other strength measures may also be used customized to the domain of application.

An implementation according to an embodiment of the present invention was applied to data on cancer occurrences in 32 counties over a 19 year period.

occurrence.txt

Grant	1	1997	2	2
SanJuan	1	1974	8	2
Bernalillo	1	1977	13	1
DonaAna	1	1977	14	2
Union	1	1977	16	2
Sandoval	1	1977	11	1
Valencia	1	1977	17	1
DonaAna	1	1977	11	2
Valencia	1	1977	7	2
Bernalillo	1	1975	13	1

The file occurrence.txt contains a portion of the input occurrence data. Each line of this file specifies the

county, number of occurrences of cancer, the year of occurrence and values for the other characteristics like age group and gender of the cancer patients.

5 population.txt

Bernalillo	73	15537	1	1
Bernalillo	73	14931	1	2
Catron	73	99	1	1
Catron	73	99	1	2
Chaves	73	1888	1	1
Chaves	73	1867	1	2
Colfax	73	567	1	1
Colfax	73	526	1	1
Curry	73	2192	1	1
Curry	73	2130	1	2

location.txt

Bernalillo	66	102
Catron	8	57
Chaves	126	47
Colfax	123	162
Curry	161	79
DeBaca	132	82
DonaAna	64	11
Eddy	136	13
Grant	22	25
Guadalupe	129	97

10

The files location.txt and population.txt contain a portion of the characteristics of the locations and times. The file location.txt contains the Cartesian coordinates defining the location of each county being analyzed. Each line of the file population.txt contains a county name, year, and population for each set of values of the other characteristics (e.g., age group and gender). The cluster

YOR92004008US1 (8728-673)

detected with the square pyramid as chosen convex container shape is shown in Figure 6. This cluster extends from year 1982 to the year 1989. There are 292 cases included in this cluster, which is significantly larger than the 211.57
5 expected if there were no clusters using the Poisson model and indirect standardization. The results in Figure 6 demonstrate how the system and method are able to capture characteristics like growth and movement in the phenomenon (e.g., cancer occurrences) that would not be possible with
10 the prior art.

Referring to Figure 7, a method according to an embodiment of the present disclosure takes as input data on occurrences of a phenomenon at locations and times and data on characteristics of the locations and times 701. The
15 occurrences of a phenomenon and characteristics of locations and times are given for points in space and time. Actual occurrences of the phenomenon are determined at the points according to the input data 702. Expected occurrences are determined in the absence of any clustering of the
20 phenomenon at the points according to the characteristics of the locations and times using a domain dependent model for the phenomenon 703. A convex container shape is selected for determining a cluster in the input data 704. The convex container shape may be selected at any time before searching

for solutions. Solutions represented as a set of points are sought that conform to the selected convex container shape for a cluster with a highest strength metric that is based on the number of occurrences included in the cluster and the input data 705.

Having described embodiments for system for detecting generalized space-time clusters, it is noted that modifications and variations can be made by persons skilled in the art in light of the above teachings. It is therefore to be understood that changes may be made in the particular embodiments of the invention disclosed which are within the scope and spirit of the invention as defined by the appended claims. Having thus described the invention with the details and particularity required by the patent laws, what is claimed and desired protected by Letters Patent is set forth in the appended claims.